DOCUMENT RESUME

ED 397 067 TM 024 412

AUTHOR Jenkins, Jerry A.
TITLE Can Quality Program Evaluation Really Take Place in Schools?
PUB DATE 15 Apr 93
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Educational Assessment; Elementary Secondary Education; *Evaluation Methods; *Program Evaluation; *Student Attitudes; Student Motivation; Teacher Education; Teacher Evaluation; Testing Problems; Test Interpretation; *Test Use
IDENTIFIERS *High Stakes Tests; Testing Effects

ABSTRACT

High-stakes assessments are those in which the results of tests or other measures can lead to decisions that may affect school administrators, teachers, and students substantially. Whether high-stakes assessment results in misleading information due to extraneous factors associated with the conditions under which the assessment occurs is explored. Among the major problems associated with high-stakes assessment is the lack of adequate training for teachers and administrators with regard to measurement issues and testing. In addition, high-stakes tests can lead to student anxiety or poor student motivation. Some assessments may not be chosen carefully, and tests may be given at inappropriate times. Teachers and administrators may focus only on scores, rather than on learning. Some solutions for the adverse affects of high-stakes testing are: (1) better teacher education in measurement concerns; (2) a reduction of the link between student achievement measures and teacher evaluation; (3) new approaches to assessment; (4) the use of multiple measures of student achievement; and (5) the promotion of student attitudes that allow them to demonstrate their educational growth. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TAC-B-42P

CAN QUALITY PROGRAM EVALUATION REALLY TAKE PLACE IN SCHOOLS?

ED 397 067

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

JERRY A. JENKINS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Jerry A. Jenkins, Ph.D.
Director
Chapter 1 Technical Assistance Center, Region C
Atlanta, Georgia

The American Educational Research Association
Annual Meeting

April 15, 1993

Educational Testing Service • 1979 Lakeside Parkway, Suite 400 • Tucker, GA 30084

BEST COPY AVAILABLE

TM024412

Introduction

Lawmakers and the public in general are showing increasing disenchantment with the quality of American education today. As a consequence, they are demanding greater accountability on the part of elementary and secondary schools. One evidence of this is the Hawkins-Stafford Amendments of 1988 (Chapter 1) to Title I of the ESEA Act of 1965. According to current regulations, Chapter 1 schools that do not achieve the minimum standard of educational growth for the state in which they are located [for a given school year] are placed into a formal, school-level "program improvement" status for the following year.[1]

In addition, Chapter 1 teachers are required to undertake an annual review of the educational progress of **each student** served by the program during the course of the year. For students who show a decline or no improvement in performance for two consecutive years as assessed by the measures detailed in the local school district's annual application, teachers must "conduct a thorough assessment of the educational needs..."[2] of those children and "...use the results of that needs assessment to modify the Chapter 1 project to meet the children's needs."[3] The measures must include "...aggregate performance and the desired outcomes described in the [local educational agency's] application,"[4] "... in both basic and advanced skills."[5]

The results of such evaluations are often published and widely advertised by local, frequently state, and sometimes national news media. This coverage almost always results in increased pressures by communities on school staffs to undertake efforts to improve their test scores as well as any other evidence they may have from other measures of student achievement. As a consequence, in many cases, higher test scores and more positive indicators of achievement are found the following school year. But that does not necessarily mean that the school program has been enhanced.

High-Stakes Assessment

Doubts often exist among community members about the accuracy of school program assessments and the quality of the evaluations conducted by school staff. These quandaries make one wonder whether evaluations involving high stakes can take place in schools. That is, can rigorous assessments be made of programs when those engaged in the data gathering are the same individuals significantly affected by the evaluation outcomes? A logical corollary question is, are there built-in biases from such assessments that typically yield unreliable results?

It might be appropriate to initially define the term "high-stakes assessment" for the purpose of this discussion. High-stakes assessment may be defined as tests or other measures whose results can lead to decisions that may substantially, at times adversely, affect school administrators, teachers, and/or students. Among possible outcomes of these decisions are (1) requirements for burdensome program improvement plan development that sometimes involves state-level staff, (2) reorganization or restructuring of schools, (3) loss or change of jobs on the part of school faculty members, (4) changes in the educational placement of individual students, and (5) loss of funds for school programs.

The issue at hand is whether high-stakes assessment results in misleading information due to extraneous factors associated with the conditions under which the assessment takes place. Test scores and other indicators of student achievement that are too high typically result in unwarranted complacency. On the other hand, spuriously low score values may lead to a flurry of activity that makes no lasting, substantive change in the school program or in individual student achievement.

It is important in this discussion to point out that although most currently popular indicators of student achievement are based upon standardized test scores, the concerns raised by this paper apply to other measures of achievement as well. Other contemporary educational assessments include terms such as other desired

outcomes, alternative assessment, authentic assessment, and portfolio assessment. But regardless of the assessment measure or measures used, it must be recognized that a problem exists in high-stakes assessment per se, before possible solutions can be contemplated.

Major Problems With High-Stakes Assessment

What are some major factors that cause biased, unreliable outcome data? Let's look at a few typical problems the Educational Testing Service (ETS) Chapter 1 Technical Assistance Center staffs have encountered in assisting state educational agencies and local schools to undertake evaluations. The following statements, although blanket statements, seem warranted, even though one or more may not apply in a particular setting.

1. Very few practicing teachers (or school administrators) have had any formal preparation in educational assessment. Thus, they have little knowledge of measurement issues. They do not understand the concepts and importance of reliability or validity, nor do they know how to construct valid alternative (performance-based) assessments. The consequences of this deficiency include non-standardized test administrations leading to inaccurate test scores.

   A further ramification of this training deficiency is a lack of understanding about how to properly interpret and use assessment data to strengthen weak educational programs.

2. Little attention is devoted by school testing staff or other administrators to assure that proper test administrations occur in the classrooms. There is little or no training given the school staffs to reinforce correct administration procedures before the tests are actually administered. Again, the consequence is poor administration and unreliable results.

3. Students have little or no training in developing test-taking skills. If the purpose of assessment is to obtain an accurate picture of a child's skill development, children should be afforded instruction on how to interpret and respond to varied test items in multiple subject areas. Items are often missed simply because children do not understand how to interpret and respond to the item. Younger children have a great deal of difficulty coping with separate answer sheets, often mis-marking answers or placing answers in the wrong place. The result of this problem is under-estimation of students' actual skill attainments.

4. Given the ramifications of low test scores, such as placement into remedial courses, grade level retention, and possible prevention from graduation, added to parental penalties, student anxiety is often extremely high. The detrimental effects of high anxiety on test performance are documented and well-known.

5. On the other hand, student motivation may be lacking. There are at least three reasons for this. First, the student's parents or peer groups feel that test scores are meaningless, thus the child makes little effort. Second, a child may have developed serious inhibitions toward assessment, based on past unsuccessful test attempts. Third, as is the case with at least one state testing program, children do not see their test scores and, therefore, have no reason to give best effort.

6. From a logistical standpoint, test batteries are often purchased on the basis of reputation, economics, and other irrelevant factors, and not on the basis of test quality and match of test content to the local curriculum. Thus, they are regularly inappropriate for local use.

7. To accommodate issues related to test scoring, tests are often administered at wrong times of the year and inappropriate test levels are

given to children. If tests are to be a meaningful source of information, they should reflect as much of the school curriculum as possible. Yet, many tests are administered in early to mid-March with roughly one-third of the school year remaining. Whether there is a match between the weight given to each of the test objectives, as determined by the number of items measuring each objective, and the content imbedded in the classroom curriculum is, at that point, "the luck of the draw."

8. Given the pressure to "produce," teachers often spend considerable time prior to each test administration teaching facts related to the test items or test objectives. This teaching to the test occurs most frequently, consciously or otherwise, when the test has been used previously. During the actual administration, teachers sometimes give children clues to answers of "hard" test items.

   When these practices do happen, the items on the test no longer represent a sampling of a larger domain, but the test becomes a body of knowledge in and of itself. Thus scores inaccurately reflect student general skill acquisition. The results of these actions are artificially high scores.[6]

9. When the score results are returned, both teachers and the administrators focus their attention almost solely on low scores, looking for possible errors in these cases only. Higher-than-normal scores are rarely examined. This exclusive focus on low scores almost certainly results in overly optimistic (positively biased) average scores.

Is There A Solution?

Although many forms of measures used in schools, standardized tests in particular, have been cited as being the root cause of inaccurate assessment, the fact

is that the problem lies primarily with the **system** and only partially with the instruments.[7] Greg Anrig, president of ETS, stated in an argument for equal educational opportunity, "We do not cure a virus by throwing away the thermometer that alerts us to the existence of a fever."[8] I would add that in order to consider the thermometer reading to be reliable, it must be properly used.

At least five changes must occur in education before school-based educational assessment and evaluation will provide reliable information.

First and foremost, there needs to be a reduction (though not elimination) of the strength of the link between measures of student achievement and evaluation of teaching performance. This move alone would substantially diminish the stakes now found in classroom assessment and could result in more valid assessments.

There is growing evidence that high-stakes testing may be more harmful than helpful. Allington and McGill-Franzen (1992) studied the outcomes of high-stakes testing in seven schools. They concluded that such use of tests "can not only obscure but even reinforce questionable educational practices-just the opposite of their original intent."[9]

The **second** change is that two courses of instruction on assessment and evaluation should be added to teacher certification requirements. One course should assure that teachers understand the nature and the importance of measurement-related criteria, such as reliability, validity, standardization of administration, relevancy, and norming. This course should also train teachers in proper test administration procedures and other measurement processes. In sum, teachers would leave this course with a clear understanding of what constitutes quality assessment of achievement.

The other course of instruction should be one that focuses on the interpretation of assessment information and the use of that information for curriculum design and

instructional improvement. This is the evaluative component that follows measurement. Simply having knowledge of students' levels of achievement is not enough. Teachers must know how to apply that knowledge in a constructive yet reasonable manner.

The **third** change should be the development of valid and reliable alternative measures to standardized tests that would still be viable for school classroom use. We in education have been guilty of using standardized test scores the way a young child might use a new hammer; i.e., everything the child sees needs hammering. That is to say, we have tried using the standardized tests for many purposes for which they were not designed. Measures which are appropriate for different purposes are badly needed.

The **fourth** change should be the recognition that multiple measures of student performance will be more informative, more accurate, and more useful than any single measure. That is, there should be a reduction on reliance on single measures, such as test scores, coupled with a push to design appropriate measures for different uses. Teachers, then, should be helped to ascertain how to construct or determine which measures are best for different uses.

Fifth, we must provide students with the necessary tools and attitudes that will allow them to demonstrate their educational growth. This may be the most difficult of all tasks.

In some cases, students will verify their knowledge and skill acquisitions through performance assessment methods. In other instances, standardized tests may be the means of proof. Test-taking skills are not inborn—they must be learned as reading and math skills are learned. This can occur only if schools develop and present deliberately planned lessons on how to address and respond to the variety of test items in the different subject areas.

Conclusion

Achievement measures are indicators of all forces that impinge upon the student at the time they are acquired, not just of the success of classroom instruction. To determine the actual extent of academic achievement requires the best measurement possible. At present, there are many factors that preclude valid assessment. For that to occur, changes need to take place to improve all aspects of the assessment process. Until we make these changes, we should not expect nor are we likely to get qualitative evaluations of programs from schools.

NOTES

1. Section 200.37(a)(2)(i) and §200.38(b), Federal Register, USED, May 19, 1989.
2. Federal Register, §200.38(d)(3).
3. Ibid. (4).
4. Ibid. (a)(1)(i).
5. Federal Register, §200.35(1)(i).
6. For additional discussion of this topic, see "Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution," Educational Leadership, Washington, D.C, June-July 1991, pp. 2-7.
7. See for example, "Uses and Abuses of Achievement Test Scores," Educational Measurement: Issues and Practice, Washington, D.C., Summer 1992, pp. 9-15.
8. "Educational Standards, Testing, and Equity," Phi Delta Kappan, May 1985, p. 624.
9. "Does High-Stakes Testing Improve School Effectiveness?", Richard L. Allington and Anne McGill-Franzen, ERS Spectrum, Vol. 10, No. 2, Spring 1992, pp. 3-12.

The opinions expressed in this paper are solely those of the author and do not necessarily represent the views or policies of the United States Department of Education nor of Educational Testing Service.